Jurnal Teknologi

Full paper

Analysis for Soluble Solid Contents in Pineapples using NIR Spectroscopy

Herlina Abdul Rahim*, Chia Kim Seng, Ruzairi Abdul Rahim

Protom-i Research Group, Infocomm Research Alliance, Control and Mechatronic Engineering Department, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding author: herlina@fke.utm.my

Article history

Received :4 Februari 2014
Received in revised form :
6 April 2014
Accepted :4 May 2014

Graphical abstract

Abstract

This research investigates the use of predictive models and a low cost spectroscopy in non-invasive Soluble Solids Content (SSC) assessment. The challenge is to model complex and high-dimensional spectral data. Results indicate that the SSC prediction of pineapples harvested on different days using the proposed approach is promising. The second study evaluates common pre-processing practices. Findings indicate that the best accuracy would be achieved when visible spectrum was excluded, second order Savitzky-Golay derivative with the optimal filter length was used, and the absorbance transformation was avoided.

Keywords: VIS-SWNIR; soluble solid contents; pineapples; principal component regression

1.0 INTRODUCTION

In the NIR spectroscopic analysis, Principal Component Analysis (PCA) has been widely implemented to compress a high-dimensional spectrum into few variables as the predictors of Artificial Neural Network (ANN) to overcome collinearity and redundancy problems [1, 2]. However, the classical statistical estimators, such as mean and standard deviation, can be easily affected by outlying variables; thus classical estimators of PC [3]. This implies that the use of a robust PCA that minimizes the effects of potential outlying observations in spectral data might be useful to improve the accuracy and robustness of a predictive model.

Next, common spectral pre-processing by means of absorbance transformation has been widely applied to improve the relationship between the spectral data and components of interest [3-5]. However, this transformation may not essentially improve the multivariate linearity [6]. The second order Savitzky-Golay (SG) derivative pre-processing, on the other hand, is capable of removing both baseline and slope effects, and attenuating the effects of high frequency noise, simultaneously. However, the use of second order SG derivative has seldom been reported in traceable literature compared to zero order and first order SG derivative. This could probably due to the lack of studies on the use of tunable parameters in SG derivative [7]. An investigation on the necessity of the absorbance transformation, and the effect of SG derivatives against different settings might offer explanations for these concerns.

2.0 METHODOLOGY

Generally, the three vital steps in the spectroscopic analysis are data pre-processing, model design, and validation. Data pre-processing is used to enhance the quality of the acquired data by eliminating or minimizing the effects from unwanted signals, such as baseline shifts and slope effects. Model design is applied to ensure that the architecture of a predictive model is optimized during training or calibration process. Validation strategy is used to authenticate the performance of trained predictive models, and to ensure that the findings are reliable

2.1 Outlier Detection

Outlier detection was carried out to identify potential outliers in the data set of pineapple samples. For prediction analysis, outlier detection was independently implemented on each data set, in which, 48 samples per data set, by means of externally studentized residues. In order to compare different predictive models, on the other hand, all data of 192 pineapple samples were used together to identify potential outliers.

2.2 Individual Data Set

Outlier detection by means of externally studentized residuals was independently carried on individual data set for prediction analysis study in Section 4.4. Firstly, reflectance spectral data were pre-processed using first order derivative coupled with

Savitzky-Golay (SG) smoothing. The order of polynomial fit and the filter length of SG smoothing were arbitrarily set to one and 35 points or 7.38 nm. After that, robust PCA (ROBPCA) decomposed the pre-processed spectral into robust principal components (RPCs). Leave-One-Out Cross-Validation was used to determine the optimal number of predictors for Multiple Linear Regression to perform RPCR. The difference between the predicted and the measured SSC values were used to compute the externally studentized residual. The critical value of the outlier detection was based on the critical value of t-distribution. With confidence level of 80% and degree of freedom of 40, which is an approximation of 48 samples per data set, the critical value for a two-tails t-distribution is 1.303 [8,9]

2.3 All Data Sets

To avoid potential masking effects during outlier detection, Monte Carlo Cross-Validation (MCCV) with 1000 iterations and 50% leave-out was implemented [10]. Firstly, the interval of spectral data was fixed to one nm using the methodology. After that, visible and shortwave near infrared (VIS-SWNIR) reflectance spectra were pre-processed by second order Savitzky-Golay (SG) derivative. The filter length of the SG derivative was arbitrarily fixed to 17 nm. Next, Adaptive Linear Neuron (ADALINE) was used to compute the Root Mean Square Error of Cross-Validation (RMSECV) of retained samples when the critical value was varied. The learning rate of ADALINE was intuitively set to 0.01. The effect of adaptation cycle was investigated. Plots of the predicted SSC values versus the measured SSC values, and the standardized residue values versus predicted SSC were used for visualizing the potential outliers. The optimal critical value should retain sufficient samples for analysis with satisfactory performance.

2.4 Interval Correction

The low cost portable visible and shortwave near infrared (VIS-SWNIR) spectroscopy uses a charge coupled device (CCD) linear image sensor (TCD1304AP, Toshiba, Japan) that contains 3648 detector elements to provide a spectrum between 650 and 1318 nm. However, the CCD sensor is unable to measure spectrum above 1100nm, and substantial noise existed at the beginning and the end of the acquired spectrum. Therefore, only 1741 wavelengths that are from 60 to 1800 detector elements were retained, giving spectral information from approximately 662 to 1005 nm.

The interval between two adjacent wavelengths is inconsistent, as depicted in Figure 1. For instance, interval values between four consecutive wavelengths of 668.76, 668.97, 669.17, and 669.38 nm are 0.21, 0.20, 0.21, and 0.21 nm, respectively, and decreases from 0.21 to 0.18nm along the spectrum. This may be due to analogue-to-digital conversion and the decrease of grating efficiency of the spectroscopy.

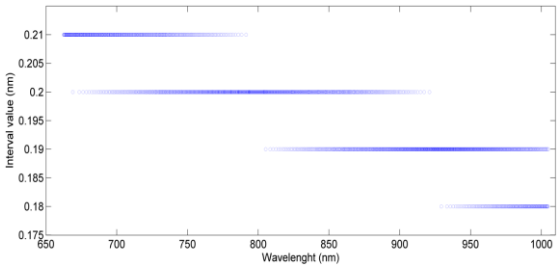

Figure 1 The interval value versus the wavelengths

A simple averaging approach was applied to fix the interval of spectral data to one nanometer. Firstly, the wavelength values of a spectrum were rounded to the nearest integer. For instance, five consecutive wavelengths of 668.76, 668.97, 669.17, 669.38, and 669.59 nm were rounded to 669, 669, 669, 669, and 670 nm, respectively. Consequently, the number of wavelengths in a spectrum that ranged from 662 to 1005 nm with an interval of one nm was 344. The reflectance value for each wavelength was then taken as the average of the reflectance values of wavelengths that have a same nearest integer. For example, the reflectance value of 669 nm was the average of the reflectance values of 668.76, 668.97, 669.17, and 669.38 nm. This averaging approach was implemented by Ventura [16] to overcome

■3.0 RESULTS AND DISCUSSION

The finding of the research is delivers of the result of outlier detection described in Section 3.1. RMSEC and RMSEP are used to discuss the calibration and predictive accuracies of a predictive model, respectively.

3.1 Outlier Detection

The findings of outlier detection on pineapple data set are presented in the following sub-sections. Section 5.3.1 delivers the findings of outlier detection on data sets of Day 1, Day 2, Day 3, and Day 4, individually.

3.2 Individual Data Set

Figure 2 illustrates the acquired spectral data that were pre-processed using the first order derivative coupled with Savitzky-Golay (SG) smoothing. This pre-processing approach substantially eliminated the baseline shift that existed in the acquired spectral data as that illustrated in Figure 2.

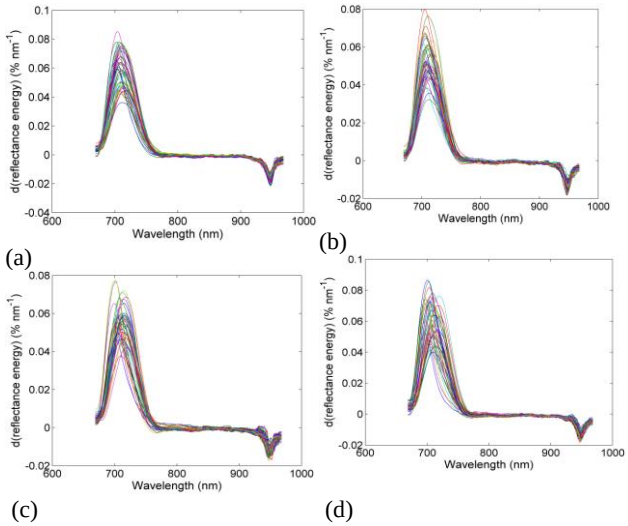

Figure 2 The acquired reflectance spectral data that were pre-processed by using the first order derivative coupled with SG smoothing, from (a) Day 1, (b) Day 2, (c) Day 3, and (d) Day 4

Figure 3 suggests that the optimal predictor number of RPCR in each data set was two. This is because Root Mean Square Error of Cross-Validation (RMSECV) values were not or marginally improved when more than two predictors were used.

Moreover, the exclusion of those low variance predictors can avoid over-fitting problems, especially for the analysis of a small data set of 48 samples.

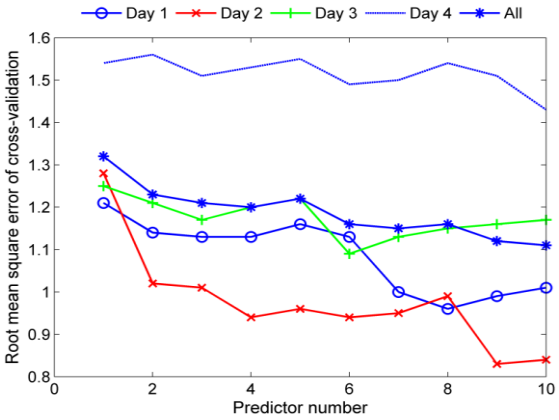

Figure 3 Rmsecv versus the number of predictors

Figure 4 illustrates that the potential outliers and inliers in each data set based on externally studentized residues, in which, only the first two RPCs were used as the predictors of MLR to perform RPCR.

Table 1 summarizes the measured SSC values of the four data sets that measured from different days after excluding potential outliers based on externally studentized residual. The descriptive statistics of the measured SSC values from different data sets were different because the pineapples were harvested and randomly chosen from their populations on the different days. The measured SSC values of the retained 157 pineapple samples, which were acquired on the four different days, were approximately normally distributed around a mean of 11.03°Brix with a standard deviation of 1.25°Brix, as illustrated in Figure 5.

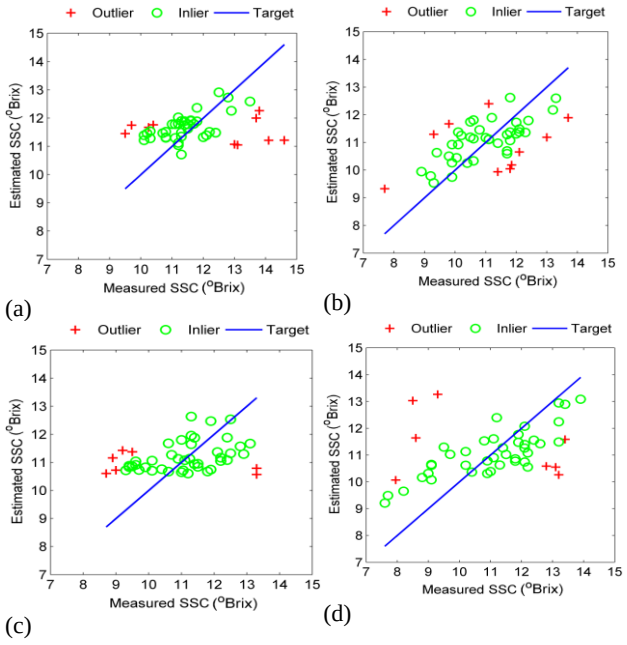

Figure 4 Outlier detection by means of externally studentized residual on individual data sets: (a) Day 1, (b) Day 2, (c) Day 3, and (d) Day 4

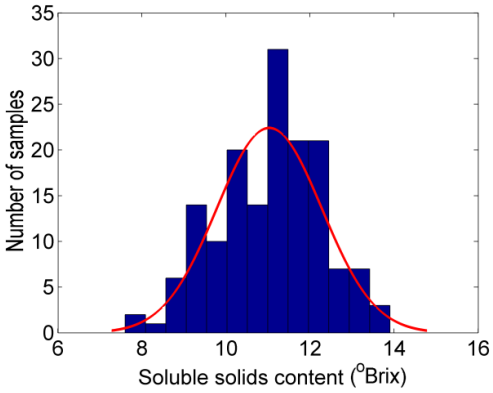

Figure 5 Histogram of the measured SSC values of the 157 pineapple samples

Table 1 Descriptive statistical of the SSC of the retained pineapple samples

Data set	Number of samples	SSC (°Brix)			Standard deviation
		Min	Max	Mean	
Day 1	38	9.5	13.8	1.30	0.98
Day 2	38	8.9	13.3	0.96	1.10
Day 3	41	8.7	13.1	0.82	1.22
Day 4	40	7.6	13.9	1.05	1.60
Total	157	7.6	13.9	1.03	1.25

3.3 All Data Sets

Figure 6 illustrates the shortwave near infrared (SWNIR) reflectance spectra that were pre-processed by second order Savitzky-Golay (SG) derivative. After the spectral data had been pre-processed by second order SG derivative, the positive and negative peaks that appear in Figure 6 (the first order derivative) were approximated to zero. The peaks of the second order SG derivative spectra, on the other hand, denote the maximum change of the slopes along a spectrum.

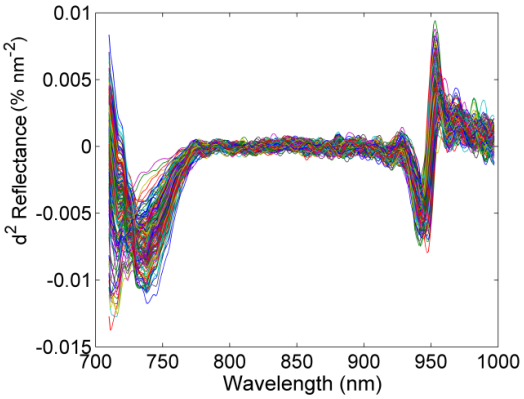

Figure 6 SWNIR reflectance spectra after second order Savitzky-Golay derivative

Figure 7 depicts that the lower the critical value, the smaller the RMSECV with fewer retained samples. This is possible because fewer retained samples reduce the chance of including bad samples that degraded the performance of a predictive model. When the critical value was 1.7, ADALINE that used one

adaptation cycle and that used two adaptation cycles retained 162 retained samples, as illustrated in Figure 7(a) and Figure 7(b), respectively. However, the former achieved higher correlation coefficient of 0.77, compared with the latter, in which, correlation coefficient was 0.75. These indicate that one adaptation cycle was sufficient for ADALINE to reach its optimal performance when its learning rate was 0.01.

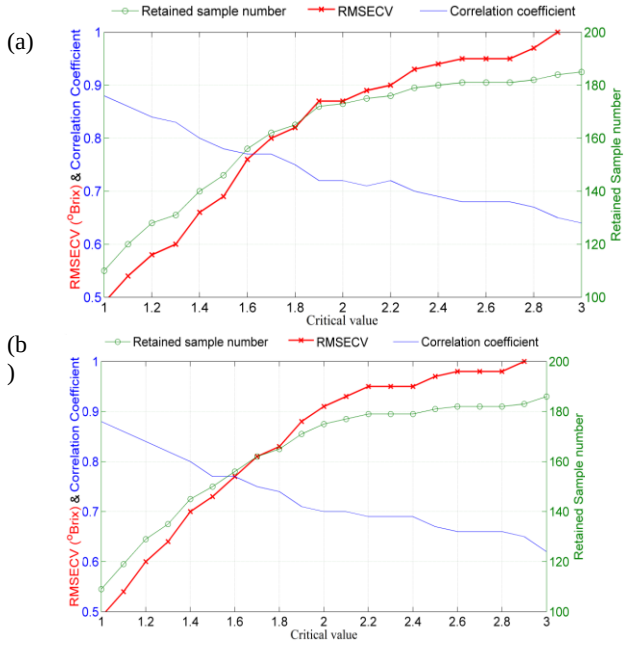

Figure 7 Relationships among validation performance, retained samples, and critical value for ADALINE: (a) that used one adaptation cycle, and (b) that used two adaptation cycles

According to the SSC residues distribution of the 192 pineapple samples in Figure 8(a), the nonlinearity information in the data is negligible. This is because the residuals were fairly normally distributed at the mean of 0.0056°Brix with standard deviation of 1.1607°Brix. Nonetheless, few extreme observations in Figure 8(b) indicate that they were potential outliers. The relationship between predicted SSC and measured SSC values appears to be substantially improved when the potential outliers were excluded, as illustrated in Figure 8(c). Thus, a total of 162 samples were retained with the critical value of 1.7.

It is worthy to highlight that although the samples used in the present section were the same as the previous, that is, 192 pineapple samples, outlier detection used in the present section was in different manner. First, outlier detection was carried out by considering all the 192 samples, instead of the 48 samples on each day individually in the previous section. Hence, it should not be an unexpected result when five more samples were retained in the present section, compared to the previous section of 157 retained samples. Another possible reason is that the use of Monte Carlo Cross-Validation (MCCV) rectified the inconsistency of Leave-One-Out Cross-Validation (LOOCV) without the requirement of a "balanced" collection of subsets [11]

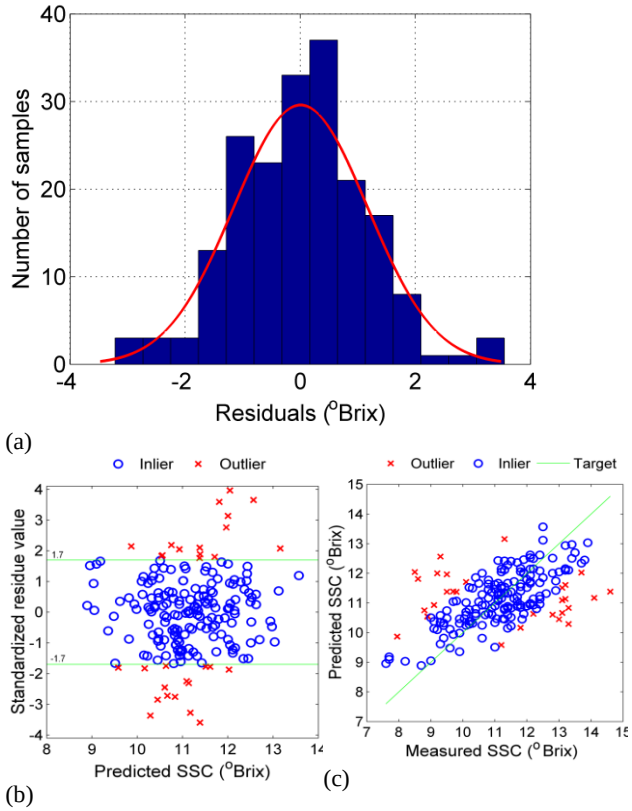

Figure 8 Diagnosis plots when critical value is 1.7: (a) the histogram of residuals of 192 pineapple samples, (b) standardized residue values versus the predicted SSC, and (c) the predicted SSC versus the measured SSC values

4.0 CONCLUSION

The potential of a low cost VIS-SWNIR spectroscopy for the non-invasive SSC assessment of pineapples has been evaluated and discussed using the four data sets that acquired on different days The findings indicate that wavelengths in the vicinity of the visible spectrum (from 662 to 700 nm) did not provide unique relevant information that could be extracted by PCR to achieve better predictive accuracy in the SSC assessment of pineapples. Findings also indicate that the log (1/R) transformation degraded the performance of PCR; and the filter length had a considerable influence on the second order SG derivative. In particular, results show that PCR was able to tolerate the baseline shift and slope effects using extra PCs. However, this violets the principle of parsimonious, and thus, it is not encouraged. In short, the non-invasive SSC assessment of pineapples by means of a low cost spectrometer could be accomplished using a parsimonious PCR with 4 predictors. The best accuracy with RMSECV of 0.81°Brix and rcv of 0.75 was achieved when the visible spectrum was excluded, second order SG derivative with optimal filter length was used, and the absorbance transformation was avoided.

Acknowledgement

The authors would like to thank the Ministry of Higher Education and Universiti Teknologi Malaysia for supporting this research under R.J1300000.4L606.

References

[1] Chen, Q., Guo, Z., Zhao, J. and Ouyang, Q. 2012. Comparisons of Different Regressions Tools in Measurement of Antioxidant Activity in Green Tea Using Near Infrared Spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*. 60: 92–97.

[2] Zhao, X., He, Y. and Bao, Y. 2011. Non-Destructive Identification of the Botanical Origin of Chinese Honey Using Visible/Short Wave-Near Infrared Spectroscopy. *Sensor Letters*. 9(3): 1055–1061.

[3] Daszykowski, M., Kaczmarek, K., Vander Heyden, Y. and Walczak, B. 2007. Robust Statistics in Data Analysis-A Review: Basic Concepts. *Chemometrics and Intelligent Laboratory Systems*. 85(2): 203–219.

[4] Jamshidi, B., Minaei, S., Mohajerani, E. and Ghassemian, H. 2012. Reflectance Vis/NIR Spectroscopy for Nondestructive Taste Characterization of Valencia Oranges. *Computers and Electronics in Agriculture*. 85: 64–69.

[5] Wang, J., Nakano, K. and Ohashi, S. 2011. Nondestructive Evaluation of Jujube Quality by Visible and Near-infrared Spectroscopy. *LWT-Food Science and Technology*. 44(4): 1119–1125.

[6] Næs, T., Isaksson, T., Fearn, T. and Davies, T. 2012. Non-linearity Problems in Calibration. A User-Friendly Guide to Multivariate Calibration and Classification. *Chichester UK: NIR Publications. 93-97*; 2002.

[7] Rinnan, A., Berg, F. v. d. and Engelsen, S. B. 2009. Review of the Most Common Pre-processing Techniques for Near-infrared Spectra. *TrAC Trends in Analytical Chemistry*. 28(10): 1201–1222.

[8] Nelson, W. 2008. *Appendix A. Statistical Tables. Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*. John Wiley & Sons, Inc. 549-560.

[9] Gujarati, D. N. 2004. *Appendix D: Statistical Tables. Basic Econometrics*. Fourth Edition. New York: McGraw-Hill. 959–975.

[10] Liu, Z., Cai, W. and Shao, X. 2008. Outlier Detection in Near-infrared Spectroscopic Analysis by Using Monte Carlo Cross-validation. *Science in China Series B: Chemistry*. 51(8): 751–759.

[11] Shao, J. 1993. Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*. 88(422): 486–494.